Out-of-distribution Generalization with Causal Invariant Transformations

Ruoyu Wang[1,2*]   Mingyang Yi[1,2*]   Zhitang Chen[3]   Shengyu Zhu[3†]

[1]University of Chinese Academy of Sciences

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences

[3]Huawei Noah's Ark Lab

{wangruoyu17,yimingyang17}@mails.ucas.edu.cn, {chenzhitang2,zhushengyu}@huawei.com

Abstract

In real-world applications, it is important and desirable to learn a model that performs well on out-of-distribution (OOD) data. Recently, causality has become a powerful tool to tackle the OOD generalization problem, with the idea resting on the causal mechanism that is invariant across domains of interest. To leverage the generally unknown causal mechanism, existing works assume a linear form of causal feature or require sufficiently many and diverse training domains, which are usually restrictive in practice. In this work, we obviate these assumptions and tackle the OOD problem without explicitly recovering the causal feature. Our approach is based on transformations that modify the non-causal feature but leave the causal part unchanged, which can be either obtained from prior knowledge or learned from the training data in the multi-domain scenario. Under the setting of invariant causal mechanism, we theoretically show that if all such transformations are available, then we can learn a minimax optimal model across the domains using only single domain data. Noticing that knowing a complete set of these causal invariant transformations may be impractical, we further show that it suffices to know only a subset of these transformations. Based on the theoretical findings, a regularized training procedure is proposed to improve the OOD generalization capability. Extensive experimental results on both synthetic and real datasets verify the effectiveness of the proposed algorithm, even with only a few causal invariant transformations.

1. Introduction

The success of many machine learning algorithms with empirical risk minimization (ERM) relies on the independent and identically distributed (i.i.d.) hypothesis that training and test data originate from a common distribution. In practice, however, data in different domains or environments are often heterogeneous, due to changing circumstances, selection bias, and time-shifts in the distributions [47, 57]. Accessing data from all the domains of interest, on the other hand, is expensive or even impossible. Consequently, the problem of learning a model that generalizes well on the unseen target distributions is a practically important but also challenging task and has gained much research attention in the past decades [6, 7, 17, 49, 68].

Since data from some domains are unavailable, assumptions or prior knowledge on the unseen domains are generally required to achieve a guaranteed out-of-distribution (OOD) generalization performance. Recently, causality has become a powerful tool to tackle the OOD problem [2, 51, 55, 57]. This is based on the assumption or observation that the underlying causal mechanism is invariant in general, even though the data distributions may vary with domains. It has been shown that a model would perform well across different domains in the minimax sense if such a causal mechanism is indeed captured.

To capture the invariant causal mechanism, existing works have assumed a particular form of the causal diagram [25,46,48,57,67], which may be restrictive in practice and is untestable from the observed data. Other works try to recover the so-called "causal feature" from the data to improve the OOD generalization performance [10, 22, 44, 55]. These works usually assume a linear form of causal feature [10, 22, 44, 57] or that there are sufficiently many and diverse training domains so that the causal feature could be identified via certain invariant properties [2, 50, 55]. In the absence of these assumptions, existing methods such as invariant risk minimization [2] can fail to capture the invariance or recover the causal feature even in simple examples [30]. In real applications like image classification, the linearity assumption may not hold, and it may be expensive or even impossible to ensure that the available domains are indeed sufficient. As such, the identifiability issue of causal feature can hardly be resolved in practice.

In this paper, we obviate the aforementioned assump-

---

*Equal contribution. Work was done during an internship at Huawei Noah's Ark Lab.

†Corresponding author.

tions and propose a new approach to learn a robust model for OOD generalization under the invariant causal mechanism assumption. We do not try to *explicitly* recover the causal feature; rather, we directly learn a model that takes advantage of the invariant properties. Our approach is based on the observation that though the explicit functional form of the causal feature is generally unknown and maybe also hard to learn, we often have some prior knowledge on the transformations that the causal feature is invariant to, i.e., transformations that modify the input data but do not change their causal feature. For example, the shape of digit in an image from `MNIST` dataset [38] can be treated as a causal feature when predicting the digit, while flipping or rotation does not change causal meanings. A detailed discussion on this issue is given in Section 3.5. We refer to these transformations as *causal invariant transformations* (CITs).

Theoretically, we prove that given complete prior knowledge of CITs, it is feasible to learn a model with OOD generalization capability using only single domain data. Specifically, we show that if all the CITs are known, then minimizing the loss over all the causally invariant transformed data, which are obtained by applying the CITs to data from the given single domain, would lead to the desired model that achieves a minimax optimality across all the domains of interest. Noticing that obtaining all CITs may be impractical, we further show that, for the purpose of OOD generalization, it suffices to know only an appropriate subset of CITs, referred to as *causal essential set* and is formally defined in Definition 2. The learned model is then shown to generalize to different domains if it is invariant to transformations in a causal essential set. This is different from existing works [59, 65] that demonstrate an improved i.i.d. generalization capability from invariance properties.

Following these theoretical results, we propose to regularize training with the discrepancy between the model outputs of the original data and their transformed versions from the CITs in the causal essential set, to enhance OOD generalization. The CITs can be viewed as data augmentation operations; in this sense, our theoretical results reveal the rationale behind data augmentation in OOD problems. Experiments on both synthetic and real-world benchmark datasets, including `PACS` [40] and `VLCS` [17], verify our theoretical findings and demonstrate the effectiveness of the proposed algorithm in terms of OOD performance. Noticeably, in some experiments, we use CycleGAN to learn the transformations between different environments, which are then used as our CITs. This is in contrast with [78] which conjectured that source-to-source transformation could provide little help to domain generation tasks in their approach.

2. Related Work

As data from some unseen domains are completely unavailable, assumptions or prior knowledge on the data distri-

butions are required to guarantee a good OOD generalization performance. We will briefly review existing domain generalization methods according to these assumptions.

Marginal Transfer Learning A branch of works assume that distributions under different domains are i.i.d. realizations from a superpopulation of distributions and augment the original feature space with the covariate distribution [6, 7]. This i.i.d. assumption on the data distribution is akin to the random effect model [8, 37] or Bayesian approach [13, 54], but may be inappropriate when the difference between domains is irregular, e.g., different styles and backgrounds in the `PACS` and `VLCS` datasets, respectively.

Robust Optimization Existing works also consider OOD data that lie close to the training distribution in terms of probability distance or divergence, e.g., Wasserstein distance [39, 64, 68, 73] or f-divergence [16, 20, 28]. They proposed to train the model via distributional robust optimization so that the model generalizes well over a set of distributions, the so-called uncertainty set [5, 60]. However, picking a suitable probability distance and range of uncertainty set in real scenarios is difficult in practice [16]. Besides, the distributions in the uncertainty set are actually the distributions of corrupted OOD data such as adversarial sample and noisy corrupted data [73], while the commonly encountered style-transformed OOD data are not included [26].

Invariant Feature Another branch of methods aim to seek a model with features whose (conditional) distributions are invariant across different domains. To this end, it was proposed to learn the feature representation by minimizing some loss functions involving domain scatter [21, 42, 49]. Here domain scatter is a quantity characterizing the dissimilarity between (conditional) distributions in different domains, as defined in [21]. [41] and [43] considered to regularize training to reduce the maximum mean discrepancy of the feature distributions of different domains and the Jensen-Shannon divergence of the feature distributions conditional on the outcome, respectively. The rationale behind these methods is to minimize a term that appears in the upper bound of the prediction error in unseen target domains [3, 4, 21]. Theoretically, the success of these methods hinges on the assumption that other terms in the upper bound are small enough [21]. However, the implication of this assumption is usually unclear and provides little guidance for the practitioner [11]. Although often not stated explicitly, the validity of these methods relies on the *covariate shift* or *label shift* assumption that is implausible if spurious correlations occur under certain domains [11, 36, 44, 76].

Invariant Causal Mechanism As in this paper, many existing works also resort to causality to study the OOD

generalization problem [1, 2, 10, 22, 25, 44, 48, 55]. In the last few years, the relationship between causality, prediction, and OOD generalization has gained increasing interest since the seminal work of [50]. The causality-based methods rest on the long-standing assumption that causal mechanism is invariant across different domains [51]. To utilize the invariant causal mechanism and hence improve OOD generalization, some works impose restrictive assumptions on the causal diagram or structural equations [25, 46, 57, 67]. Another way is through recovering the causal feature [10, 22, 55]. For example, [55] proposed to select causal variables by statistical tests for equality of distributions, and [10] leveraged some conditional independence relationships induced by the common causal mechanism assumption. It is worth noting that recovering the causal feature generally relies on restrictive assumptions, e.g., linear structural model or sufficiently many and diverse training domains [2, 10, 22, 35, 44, 50, 55]; see [56] for a further discussion on these two assumptions. Without these assumptions, existing methods such as invariant risk minimization [2] can fail to choose the right predictor even in simple examples [30]. In contrast, our approach rests on a more general causal structural model and require less training domains.

Data Augmentation Data Augmentation is an important technique to the training pipeline in deep learning [29, 34, 71, 74, 75, 79]. Commonly used methods include image rotation, cropping, Gaussian blurring, etc. With augmented data involved in training, the model generalization capability can be improved on both in-distribution [63] and out-of-distribution data [69]. Different from the above mentioned works, when domain partition is available, we apply CycleGAN [80] to learn the source-to-source translations to generate corresponding images with different styles, i.e., artificially generated casual invariant transformed data.

Next we clarify the difference between our method and the existing methods that also involve generative models to obtain augmented data [31, 77, 78]. Specifically, in [77, 78], they generate data from inexistent "novel domains", instead of the data from known domains—it was conjectured and empirically shown in [78] that source-to-source transformation could provide little help to domain generation tasks in their approach. Roughly speaking, our augmented data represent the same "causal feature" under different domains and we leverage the augmented data by contrasting them with the original ones, while existing methods add the average loss on the augmented data from inexistent domains to the objective. Also related is [31] that artificially generates the counterfactually-augmented data which modify the causal feature but keep the non-causal part. Our work is different in that we use CITs to modify the non-causal feature but keep the causal part unchanged. Moreover, in our experiments, these CITs are obtained from prior knowledge or learned from training data, without the need of *human manipulations to each training datum*.

3. OOD Generalization via Causality

In this section, we consider a general causal structural model for the OOD generalization problem. We prove that the minimax optimality of a model can be obtained via the causal feature, even if we only have access to the data from a single domain. However, as discussed in the introduction, it may be hard to recover the causal feature exactly. We hence proceed with the aid of CITs and show that a learned model can achieve the same guaranteed OOD performance.

3.1. Invariant Causal Mechanism

We begin with a formal definition of the causal structural model used in this paper. In practice, data distributions can vary across domains, but the causal mechanism usually remains unchanged [51]. We consider the following causal structural model to describe the data generating mechanism:

$$Y = m(g(X), \eta), \ \eta \perp\!\!\!\perp g(X) \text{ and } \eta \sim F, \qquad (1)$$

where X, Y are respectively the observed input and outcome, $g(X)$ denotes the causal feature, η is some random noise, and $m(\cdot, \cdot)$ represents the unknown structural function. The relationship $\eta \perp\!\!\!\perp g(X)$ means that the noise η is independent of the causal feature $g(X)$, and $\eta \sim F$ indicates that it follows a distribution F that can be unknown.

Notice that the structural model (1) imposes no assumption on the distribution of the input X. Thus, the distributions of the outcome Y can vary with X under different environments, even though Y depends on X *only* through the causal feature $g(X)$ in the data generating mechanism. Besides, there can be two correlations in the structural model, summarized as follows:

1. Although the causal feature $g(X)$ is assumed to be independent of noise η, X can correlate with noise η under a certain domain. To see this, let us consider a toy example with the observed input $X = (X_1, X_2)$. Here noise η is correlated with X_1 while η is independent of X_2. Then for the causal structural model $Y = X_2 + \eta$, we have $g(X) = X_2$ and $g(X) \perp\!\!\!\perp \eta$ while the input X is correlated to η.

2. There may exist correlations between causal feature and other spurious features, e.g., correlation between the objective shape and the image background in image classification tasks.

Unlike the invariant causal mechanism, these two correlations are supposed to vary across domains and hence are called spurious correlations [2, 70]. If not treated carefully,

the spurious correlations would deteriorate the performance of ERM-based machine learning methods and make the model perform poorly on the target domain [2, 44, 61, 62]. For instance, in an image classification task involving horse and camel, it is very likely that in the training data all the horses are on the grass while the camels are in the desert. The spurious correlation between horse/camel and the background could easily mislead the model to making predictions using the background. Consequently, the trained model would be unreliable on OOD data.

Existing works consider a similar causal mechanism to (1) while more structural assumptions are usually imposed, e.g., $g(\cdot)$ is linear and the noise is additive [22,44,50,52,55]. Our structural model (1) generalizes existing ones as we get rid of the two structural assumptions. Thus, our model constitutes a more flexible construction that is suitable to tasks in which the assumed linear or separable structural models appear implausible, e.g., the image classification task [2]. Besides, our algorithm proposed in Section 4 does not require *explicitly* learning the causal feature $g(X)$, thus avoids dealing with the identifiability issue of $g(X)$.

3.2. Generalization via Causal Feature

Throughout the rest of this paper, we focus on the distributions under structural model (1):

$$\mathcal{P} = \{P_{(X,Y)} \mid (X,Y) \sim P_{(X,Y)} \text{ under structural model (1)}\},$$

with fixed $g(\cdot)$, $m(\cdot, \cdot)$ and F. Our goal is to train a model that generalizes well across all distributions $P_{(X,Y)} \in \mathcal{P}$ which follow the causal mechanism in structural model (1). Particularly, we aim to find a model $h^*(\cdot)$ such that

$$h^*(\cdot) \in \mathcal{H}_* := \arg\min_h \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \quad (2)$$

where $\mathcal{L}(\cdot, \cdot)$ denotes a loss function, e.g., mean squared error for regression or cross entropy for classification. A similar minimax formulation appears in many existing works for the OOD generalization problem; see, e.g., [2,9,22,44,55].

In contrast with methods based on data from sufficiently many domains [35,44,53,55,58], we next show that if $g(\cdot)$ is known, we can learn $h^*(\cdot)$ via single domain data. Let P_s be the distribution of the source domain from which the training data are collected. Denote the set of optimal models under P_s based on causal feature $g(X)$ by

$$\mathcal{H}_s = \left\{ \phi \circ g \;\middle|\; \phi(w) \in \arg\min_z \mathbb{E}_{P_s}[\mathcal{L}(z, Y) \mid g(X) = w] \right\}, \quad (3)$$

where $\circ$ is the composition of functions. In this paper, we do not distinguish two functions of w that are equal to each other "almost surely", i.e., are equal for all w except a set of probability zero. Then we have the following result.

Theorem 1. *If $P_s \in \mathcal{P}$, then $\mathcal{H}_s \subseteq \mathcal{H}_*$.*

A proof of Theorem 1 can be found in the supplementary material. Theorem 1 gives a class of models that belong to $\mathcal{H}_*$, the set of solutions to the minimax problem defined in Eq. (2). A model in $\mathcal{H}_*$ makes predictions via the causal feature $g(X)$, and can be learned using the single domain data if the form of $g(\cdot)$ is known. Theorem 1 generalizes existing results in [44, 55], in the sense that it is derived under a more general structural model and also readily includes more loss functions $\mathcal{L}(\cdot, \cdot)$ beyond the mean squared loss and cross entropy loss considered in [33].

3.3. Learning via Causal Invariant Transformation

Theorem 1 shows that it is possible to use only single domain data to learn a class of optimal models $\mathcal{H}_s$ in the minimax sense. However, such a result requires an explicit formulation of causal feature $g(X)$, which is somehow impractical [2]. On the other hand, learning causal mechanism from the observed data may face the issue of identifiability. Thus, in this section, we aim to learn a model of $\mathcal{H}_s$ without the requirement of the explicit form of $g(X)$. The idea of our method is to leverage the transformations that do not change the underlying causal feature.

Specifically, although the explicit form of $g(\cdot)$ is unknown in general, we can have prior knowledge that the causal feature should remain invariant to certain transformations $T(\cdot)$. For example, consider the horse v.s. camel problem in Section 3.1. For a given image, the shape of a horse/camel could be the causal feature that determines its category. The *exact* function w.r.t. pixels representing the shape may be hard to obtain. Nevertheless, we do know that the shape does not vary with rotation or flipping. We now formally define these transformations.

Definition 1 (Causal Invariant Transformation (CIT)). *A transformation $T(\cdot)$ is called a causal invariant transformation if $(g \circ T)(\cdot) = g(\cdot)$.*

Henceforth, $\mathcal{T}_g = \{T(\cdot) : (g \circ T)(\cdot) = g(\cdot)\}$ denotes the set consisting of all CITs. As shown in Lemma 1 in the supplementary material, the set $\mathcal{T}_g$ is quite informative for $g(\cdot)$ and hence helps resolve the OOD generalization problem according to Theorem 1. In some cases, knowing $\mathcal{T}_g$ may be equivalent to knowing the causal feature or the causal parents of the outcome, e.g., assuming linear causal mechanism. However, in applications like image classification, the causal relationships are complicated and the prior knowledge on CITs can be more accessible compared to that of causal parents, as illustrated at the end of Section 3.5.

With $\mathcal{T}_g$, the following theorem states that $\mathcal{H}_s$ is available by solving a minimax problem constructed from single domain data, even for unknown $g(\cdot)$.

Theorem 2. *If $P_s \in \mathcal{P}$, then for $\mathcal{H}_s$ defined in Eq. (3)*

$$\mathcal{H}_s \subseteq \arg\min_h \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_s}[\mathcal{L}(h(T(X)), Y)]. \quad (4)$$

A proof of Theorem 2 is in the supplementary material. If the minimax problem (4) has a unique minimum (when, e.g., some convexity conditions on the loss function $\mathcal{L}(\cdot, \cdot)$ hold), Theorem 2 implies that the model performs uniformly well over the transformed data obtained from the transformations in $\mathcal{T}_g$ can generalize to distributions in $\mathcal{P}$.

Let $\mathcal{P}_{\mathrm{aug}} = \{P_{(X',Y)} \mid (X,Y) \sim P_{\mathrm{s}},\, X' = T(X),\, T \in \mathcal{T}_g\}$, then we can rewrite the minimax problem in (4) as

$$\min_h \sup_{P \in \mathcal{P}_{\mathrm{aug}}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \qquad (5)$$

Problem (5) has a similar form to problem (2). Recalling the structural model (1), it can be verified that $\mathcal{P}_{\mathrm{aug}}$ is a subset of $\mathcal{P}$. We then have the following two remarks:

1. $\mathcal{P}_{\mathrm{aug}}$ can be a proper subset of $\mathcal{P}$. Thus, the supremum taken in (5) is more tractable compared with that in (2), as we require less information of $\mathcal{P}$. To see this, suppose that $(X, \eta) \sim P_{(X,\eta)} = P_X \times F$ for a distribution P_X and $P_{\mathrm{s}} = P_{(X,m(g(X),\eta))}$. Then for any $P \in \mathcal{P}_{\mathrm{aug}}$, we have $Y \perp\!\!\!\perp X \mid g(X)$ if $(X,Y) \sim P$. However, there can exist $P' \in \mathcal{P}$ so that X is correlated with η and hence the conditional independence no longer holds. That is, P' lies in $\mathcal{P}$ but not in $\mathcal{P}_{\mathrm{aug}}$.

2. In the horse v.s. camel example in Section 3.1, the spurious correlations lead to misleading supervision. The set $\mathcal{P}_{\mathrm{aug}}$, on the other hand, is likely to contain distributions that do not have these spurious correlations or even entail opposite correlations. Thus, the model that overfits spurious correlations can not generalize well on these distributions, and can not be a solution to problem (5). For example, the data from some $P \in \mathcal{P}_{\mathrm{aug}}$ can have most horses on the desert while most camels are on grass. Thus, the model that overfits the spurious correlation between animal and background does not perform well on this distribution.

Although Theorem 2 provides a way to learn a model with guaranteed OOD generalization, it may be computationally hard to calculate the supremum over $\mathcal{T}_g$ when it contains plenty of or possibly infinite transformations. Take image classification tasks for example. Suppose that $\mathcal{T}_g$ contains rotations of θ degree, with $\theta = 1, \ldots, 360$. Computing the loss over a total of 360 transformations is computationally expensive. Thus, it is natural to ask a question: can we substitute $\mathcal{T}_g$ in (4) with a proper subset?

3.4. Learning via Causal Essential Set

In this section, we positively answer the question at the end of Section 3.3. We show that it is sufficient to use only a subset of $\mathcal{T}_g$, referred to as *causal essential set*. Next, we first give a formal definition of causal essential set and then prove that it is indeed the desired subset.

Definition 2 (Causal Essential Set). *For $\mathcal{I}_g \subseteq \mathcal{T}_g$, $\mathcal{I}_g$ is a causal essential set if for all x_1, x_2 satisfying $g(x_1) = g(x_2)$, there are finite transformations $T_1(\cdot), \cdots, T_K(\cdot) \in \mathcal{I}_g$ such that $(T_1 \circ \cdots \circ T_K)(x_1) = x_2$.*

Clearly, there may be multiple causal essential sets, e.g., $\mathcal{T}_g$ itself is a causal essential set. In most cases, we believe that there exists $\mathcal{I}_g$ that is a proper subset of $\mathcal{T}_g$. For example, rotation with one degree itself forms a causal essential set if $\mathcal{T}_g$ is the set of rotations with $\theta = 1, \ldots, 360$, degrees.

The next theorem indicates that the prior knowledge on any such causal essential set is sufficient to achieve a guaranteed OOD generalization, using only single domain data. A proof is provided in the supplementary material.

Theorem 3. *If $P_{\mathrm{s}} \in \mathcal{P}$, then for any $\mathcal{I}_g$ that is a causal essential set of $g(\cdot)$ and $\mathcal{H}_{\mathrm{s}}$ defined in (3)*

$$\begin{aligned} \mathcal{H}_{\mathrm{s}} &= \arg\min_h \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)], \\ &\text{subject to } h(\cdot) = (h \circ T)(\cdot),\ \forall\, T(\cdot) \in \mathcal{I}_g. \end{aligned} \qquad (6)$$

Compared with Theorem 2, "$\subseteq$" related to $\mathcal{H}_{\mathrm{s}}$ in (4) is replaced by "$=$" in this theorem, which is a stronger theoretical result. Thus, one can also readily obtain the model that generalizes well on OOD data by minimizing the loss w.r.t. any data distribution induced by structural model (1), but require less CITs. In certain cases, the structure of a causal essential set is simple and is possible to be identified. Due to space limit, this is illustrated by an example in Section S3.1 in the supplementary material.

3.5. Necessity of Prior Knowledge

Before ending this section, we would like to clarify that the prior knowledge on causal invariant transformations can be more manageable, compared with prior knowledge required by many existing approaches.

First, assumption or prior knowledge is necessary to get guaranteed causal results from observational data. As previously introduced, some recent works assume *a prior* particular forms of causal graphs, whose correctness cannot be tested from observational data. Notice that learning causal graphs from the data, or the so-called causal discovery methods [51, 66, 81], faces the same issue. Other works that have theoretical guarantee on OOD generalization require sufficiently many and diverse training domains [2] and/or incorporate prior knowledge on causal diagrams and causal mechanisms [25, 45, 56]. Despite the valuable understanding and theoretical guarantee of OOD generalization for these causality based methods, it is not clear whether the assumed conditions or prior knowledge indeed hold in practical applications. In general, only randomized controlled experiments are the golden standard to infer the causal relationship, summarized as "no causation without manipulation" [27]. On the other hand, directly creating

Figure 1. Profiles of cows in these figures are considered as causal features, while the different backgrounds are spurious features.

"manipulated" data can lead to a better causal effect estimation. In [31], authors artificially generate counterfactually-augmented data that modify the causal feature but keep the non-causal part for each sentence. Then simple processing (e.g., directly combining the observational and the manipulated data) improves the generalization performance.

These observations together motivate us to consider: *how to generate manipulations to achieve a causal guarantee at a lower human cost?* For example, one need not apply human manipulations to each training datum. This consideration leads us to the CITs, on which we often have some prior knowledge. Many augmentation techniques can be used as CITs, such as rotation and Gaussian blurring. When domain partition is available, e.g., in the multidomain learning setting [2, 50], the CITs can be learned from the data. For example, we employ GAN or other generative models to synthesize "manipulated" data in Section 5.2. We believe that such prior knowledge on CITs is more accessible than artificially manipulated data, and is manageable in practical applications.

Finally, we clarify that prior knowledge on CITs can also be more accessible than that of causal feature in many tasks, as illustrated by Figure 1 from [18]. In Figure 1, profiles of cows (highlighted by green lines) are considered as causal feature, and a prediction model *only* based on such causal feature can generalize well to images with different backgrounds. However, we do not know which pixels or what function of pixels represent the profile; indeed, the profile depends on different pixels in the first and the second images in Figure 1. In contrast, we readily know that changing the background of the image, e.g., transforming the second image to the third, will not affect the causal feature. This can be treated as our prior knowledge on CITs.

4. Algorithm

We now propose an algorithm based on the previous analysis w.r.t. CITs. Let $D(\cdot, \cdot)$ denote some measure of discrepancy satisfying $D(v_1, v_2) = 0$ if $v_1 = v_2$ and $D(v_1, v_2) > 0$ otherwise. Then for any model $h(\cdot)$ and transformation $T(\cdot)$, $\mathbb{E}_{P_s}[D(h(X), h(T(X)))] = 0$ implies $h(\cdot) = h(T(\cdot))$ almost surely. Together with Theorem 3, we consider the following formulation

$$\min_h \ \mathbb{E}_{P_s}[\mathcal{L}(h(X), Y)],$$

$$\text{subject to } \mathbb{E}_{P_s}\left[\sup_{T \in \mathcal{I}_g} D(h(X), h(T(X)))\right] = 0,$$

Algorithm 1 Regularized training with Invariance on Causal Essential set (RICE).

Input: Training set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, batch size S, learning rate η, training iterations N, model $h_\beta(\cdot)$ with parameter β, initialized parameter β_0, regularization constant λ_0, causal essential set $\mathcal{I}_g$, and discrepancy measure $D(\cdot, \cdot)$.

1: **for** $i = 1, \ldots, n$ **do**
2:      Generate transformed samples $\{T(x_i)\}_{T \in \mathcal{I}_g}$.
3: **end for**
4: **for** $t = 0, \cdots, N$ **do**
5:      Randomly sample a mini-batch $\mathcal{S} = \{(x_{t_1}, y_{t_1}), \cdots, (x_{t_S}, y_{t_S})\}$ from training set.
6:      Fetch the transformed samples $\{T(x_{t_1})\}_{T \in \mathcal{I}_g}, \cdots, \{T(x_{t_S})\}_{T \in \mathcal{I}_g}$.
7:      Update model parameters via first-order method e.g., stochastic gradient descent:

$$\beta_{t+1} = \beta_t - \frac{\eta}{S} \sum_{i=1}^{S} \nabla_\beta \mathcal{L}(h_\beta(x_{t_i}), y_{t_i})\Big|_{\beta=\beta_t} +$$

$$\eta \nabla_\beta \left\{ \frac{\lambda_0}{S} \sum_{i=1}^{S} \sup_{T \in \mathcal{I}_g} D(h_\beta(x_{t_i}), h_\beta(T(x_{t_i}))) \right\}\Big|_{\beta=\beta_t}.$$

8: **end for**

---

where $\mathcal{I}_g$ is a causal essential set. To obviate the difficulty of solving a constrained optimization problem, we further consider to minimize a regularized formulation

$$\mathbb{E}_{P_s}[\mathcal{L}(h(X), Y)] + \lambda_0 \mathbb{E}_{P_s}\left[\sup_{T \in \mathcal{I}_g}[D(h(X), h(T(X)))]\right] \tag{7}$$

with a given regularization constant $\lambda_0 > 0$. Supposing that we have training samples $\{(x_i, y_i)\}_{i=1}^n$, then we propose to minimize the empirical counterpart of (7)

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(h(x_i), y_i) + \frac{\lambda_0}{n} \sum_{i=1}^{n} \sup_{T \in \mathcal{I}_g}[D(h(x_i), h(T(x_i)))].$$

We then propose Algorithm 1, Regularized training with Invariance on Causal Essential set (RICE), to solve the above problem, where the update step in line 7 can be substituted by other optimization algorithms, e.g., Adam [32].

Note that obtaining a complete causal essential set may also be hard in many applications. Nonetheless, we usually have or can learn certain transformations with the desired causal invariance. We will empirically show that the proposed RICE enables an improved OOD generalization, even with a set of only a few CITs. In this case, we can simply replace $\mathcal{I}_g$ with this set in Algorithm 1.

5. Experiments

In this section, we empirically evaluate the efficacy of the proposed algorithm RICE on real-world datasets. We train

the model using data from some of the available domains and evaluate the performance on the data from the rest domains that are not used in training. As suggested in [72], the OOD data can be classified into two categories, namely, data with *correlation shift* or with *diversity shift*. Empirical results show that RICE can handle both kinds of OOD data. Due to space limit, part of results, including a toy experiment of synthetic data mentioned in Section 3.4 and ablation studies, are given in the supplementary material.[1]

5.1. Breaking Spurious Correlation

As we have discussed in Section 3.1, the spurious correlation in the data may mislead the model to wrong predictions on OOD data, resulting in correlation shift. In this section, we empirically verify that RICE in Algorithm 1 can obviate overfitting such spurious correlations.

Data We use the colored MNIST (C-MNIST) dataset from [15]. As in [15], we vary the colors of both foreground and background of an image.

The original MNIST dataset consists of handwritten digits from ten categories, namely, 0 to 9. To construct a training set of C-MNIST, we pick two colors for the foreground of the images in a given category, and then randomly replace the foreground color with one of the two colors assigned to the category. The background color of each image is handled similarly. For the test set, we randomly assign colors to the foreground and background of each image from the MNIST test set, regardless of its category. Some images from the generated C-MNIST dataset are visualized in Figure S3 in the supplementary material. The construction introduces a spurious correlation between category and color in the training set, but not in the test set. In the following, we will show that the proposed method RICE will not be affected by this spurious correlation.

Setup Our model is a five-layer convolution neural network as in [15]. For RICE, we choose the cross entropy loss for $\mathcal{L}(\cdot, \cdot)$ and the ℓ_2-distance for $D(\cdot, \cdot)$. The model is updated by Adam [32], and other hyperparameters are given in the supplementary material. For a handwritten digit, it is known that the shape of its foreground, rather than the color of either foreground or background, determines its category. Thus, transforming the image background with a color (e.g., black) and its foreground with another color (e.g., white) would be a desired CIT. In our experiment, we simply use the original MNIST images as the transformed data to show the effectiveness of the proposed method.

As we use the original MNIST dataset in training, the training data can be seen from two domains, i.e., the original MNIST and the C-MNIST datasets. As such, we com-

---

Table 1. Accuracy (%) on the C-MNIST test set.

Dataset	ERM	Mixup	MTL	GroupDRO	DANN	IRM	RICE(OURS)
C-MNIST	13.3	17.5	14.7	14.1	28.1	15.8	**96.9**

pare RICE with several widely used domain generalization algorithms using the same training data, including empirical risk minimization (ERM), ERM with Mixup [75, Mixup], marginal transfer learning [6, MTL], group distributionally robust optimization [58, GroupDRO], domain-adversarial neural networks [19, DANN] and invariant risk minimization [2, IRM]. See the supplementary material for a further introduction of these algorithms. For these baseline algorithms, the hyperparameters are adopted from [23]. We remark that the same training data, i.e., MNIST and C-MNIST, are used in our method and baseline methods

Main Results The empirical results are reported in Table 1. We observe that only the proposed algorithm RICE works well on the OOD data in this experiment. We speculate that this is because, for the baseline algorithms, the misleading supervision signal from the color is memorized by the models, even though the original MNIST images are also included in the training set. However, for RICE, the regularizer penalizes the discrepancy between the model outputs of the colored images and the corresponding MNIST versions, which makes the model insensible to the spurious correlation but more dependent on the invariant causal feature.

5.2. Generalizing to Unseen Domains

In this section, we conduct experiments on two benchmark datasets, PACS and VLCS, which are commonly used in domain generalization. The two datasets correspond to the diversity shift that we have mentioned. As in other related works in domain generalization [23, 76, 77], the domain labels are known in the training set.

Data PACS is an image classification dataset consisting of data from four domains of different styles, i.e., {art, cartoon, photo, sketch}, with seven different categories in each domain. VLCS is a dataset comprised of four photographic domains: {VOC2007, LabelMe, Caltech101, SUN09}, and each domain contains five different categories.

Setup As in [23], we use the model ResNet50 [24] pre-trained on ImageNet [14] as the backbone model, and fine-tune the model with different baseline methods. For RICE, the model is trained by Adam and the used hyperparameters are provided in the supplementary material. To implement RICE, we need to generate casually invariant transformed data. In the PACS dataset, each domain represents a style of images, e.g., photo or art. Since varying the style of an image does not change its category, we construct the transformations that modify image styles as CITs.
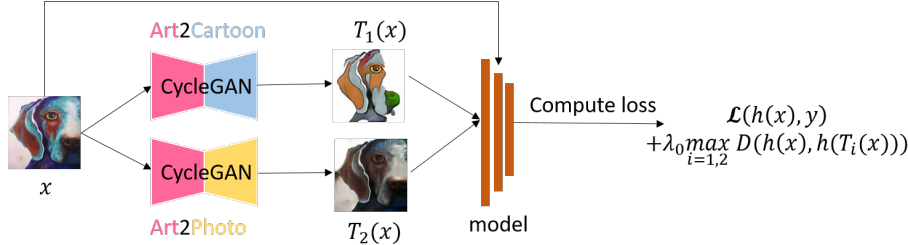
Figure 2. The proposed algorithm RICE on the PACS dataset. The training data are from domains {art, cartoon, photo}, and we want the model to perform well on the sketch data. This figure describes the training procedure of RICE for a training image from the art domain.

To this end, we use CycleGAN [80] to learn the transformations for each pair of domains in the training set, and then implement RICE using the trained CycleGAN models. In VLCS, the photographic of the image plays a similar role to the style in PACS, and we also apply CycleGAN to learning the transformations. Other generative models may also be used, e.g., StarGAN [12], for data from numerous domains.

The procedure of RICE is summarized in Figure 2. We also compare the proposed RICE with other commonly used domain generalization algorithms, as in the previous experiment. For better results and a fair comparison with baseline methods, we also provide an ablation study with single domain training data in the supplementary material.

Main Results The experimental results on PACS and VLCS are summarized in Tables 2 and 3, respectively. The results of baseline methods are from [23]. The proposed RICE exhibits better OOD generalization capability compared with the baseline methods on the PACS and VLCS datasets, in terms of the average and particularly the worst-case test accuracies. Here we provide an intuitive explanation to the performance of RICE using PACS as an example. From Figure S4 in the supplementary material, we can see that the trained CycleGAN model is likely to introduce spurious correlation related to domains, and models which capture the spurious correlation would be penalized in RICE. RICE can achieve an improved performance because it seeks to make predictions on top of the causal feature (e.g., shape of the object in the images), rather than the spurious feature related to the domains (e.g., style for PACS). Moreover, as also seen from Figure S4, some generated images from CycleGAN are indeed not similar to the original images and are also blurring, which indicates that our RICE is robust to the quality of generated data.

Finally, we verify whether the improved performance of RICE originates solely from the augmented data generated by CycleGAN. We conduct ERM training with these augmented data on PACS. The test accuracies on domains P, A, C, S are respectively 96.2, 84.9, 81.2, 80.5, which are worse than those of RICE. We believe that such a performance, which is consistent with previous observations in [78], demonstrates that the regularizer that compares dif-

Table 2. Test accuracy (%) of ResNet50 on the PACS dataset.

Method	P	A	C	S	Avg	Min
ERM	97.2	84.7	80.8	79.3	85.5	79.3
Mixup	97.6	86.1	78.9	75.8	84.6	75.8
MTL	96.4	87.5	77.1	77.3	84.6	77.1
GroupDRO	96.7	83.5	79.1	78.3	84.4	78.3
DANN	97.3	86.4	77.4	73.5	83.6	73.5
IRM	96.7	84.8	76.4	76.1	83.5	76.1
RICE (OURS)	96.8	87.8	84.3	84.7	**88.4**	**84.3**

Table 3. Test accuracy (%) of ResNet50 on the VLCS dataset.

Method	V	L	C	S	Avg	Min
ERM	74.6	64.3	97.7	73.4	77.5	64.3
Mixup	76.1	63.4	98.4	72.9	77.7	63.4
MTL	75.3	64.3	97.8	71.5	77.2	64.3
GroupDRO	76.7	63.4	97.3	69.5	76.7	63.4
DANN	77.2	65.1	99.0	73.1	78.6	65.1
IRM	77.3	64.9	98.6	73.4	78.5	64.9
RICE (OURS)	75.1	69.2	98.3	74.6	**79.3**	**69.2**

ferent augmentations with the original data is critical to capture the invariance and to improve the model robustness.

6. Concluding Remarks

In this paper, we theoretically show that knowledge of the CITs makes it feasible to learn an OOD generalized model via single domain data. The CITs can be either obtained from prior knowledge or learned from training data, without the need of human manipulations to each training datum. Inspired by our theoretical findings, we propose RICE to achieve an enhanced OOD generalization capability and the effectiveness of RICE is demonstrated empirically over various experiments.

In our experiments, we focus on image classification tasks for domain generalization. Nevertheless, our theory and the proposed algorithm can apply to other datasets once some CITs are available. For example, in natural language processing (NLP), changing the position of the adverbial or synonym substitution does not change the semantic meaning and hence can be treated as CITs. However, generating such causally invariant sentences via deep learning method is not easy. How to ease the generation and apply our RICE to NLP domain is left as a future work.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, 2020. 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Preprint arXiv:1907.02893, 2019. 1, 3, 4, 5, 6, 7

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 2

[4] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 2007. 2

[5] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. 2

[6] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22:1–55, 2021. 1, 2, 7

[7] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, 2011. 1, 2

[8] Howard D Bondell, Arun Krishna, and Sujit K Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010. 2

[9] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404 – 426, 2020. 4

[10] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, 2020. 1, 3

[11] Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. Preprint arXiv:2010.15764, 2020. 2

[12] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 8

[13] JJ Deely and DV Lindley. Bayes empirical bayes. *Journal of the American Statistical Association*, 76(376):833–841, 1981. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 7

[15] Richard Socher Devansh Arpit, Caiming Xiong. Predicting with high correlation features. Preprint arXiv:1910.00164, 2019. 7

[16] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. 2

[17] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *IEEE International Conference on Computer Vision*, 2013. 1, 2

[18] Cloudera Fast Forward. Causality for machine learning, 2021. https://ff13.fastforwardlabs.com/. 6

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 7

[20] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. Preprint arXiv:1712.06050, 2020. 2

[21] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2016. 2

[22] Jaime Roquero Gimenez and James Zou. Identifying invariant factors across multiple environments with kl regression. Preprint arXiv:2002.08341, 2021. 1, 3, 4

[23] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 7, 8

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[25] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021. 1, 3, 5

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision*, 2021. 2

[27] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. 5

[28] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, 2018. 2

[29] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. Preprint arXiv:1909.10351, 2019. 3

[30] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4069–4077. PMLR, 13–15 Apr 2021. 1, 3

[31] Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020. 3, 6

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6, 7

[33] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. Preprint arXiv:2008.01883, 2020. 4

[34] Alex Krizhevsky, Sutskever Ilya, and Hinton Geoffrey. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55, 2014. 3

[35] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, 2021. 3, 4

[36] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018. 2

[37] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 2

[38] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[39] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, 2018. 2

[40] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, 2017. 2

[41] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[42] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Association for the Advancement of Artificial Intelligence*, 2018. 2

[43] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, 2018. 2

[44] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, 2021. 1, 2, 3, 4

[45] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. Preprint arXiv:2102.12353, 2021. 5

[46] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7313–7324. PMLR, 2021. 1, 3

[47] Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015. 1

[48] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2020. 1, 3

[49] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 2013. 1, 2

[50] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016. 1, 3, 4, 6

[51] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 1, 3, 5

[52] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019. 4

[53] Qi Qian, Shenghuo Zhu, Jiasheng Tang, Rong Jin, Baigui Sun, and Hao Li. Robust optimization over multiple domains. In *Association for the Advancement of Artificial Intelligence*, 2019. 4

[54] Kolyan Ray and Aad van der Vaart. Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020. 2

[55] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. 1, 3, 4

[56] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020. 3, 5

[57] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021. 1, 3

[58] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 4, 7

[59] Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. Preprint arXiv:1910.06552, 2019. 2

[60] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017. 2

[61] Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. 4

[62] Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. Stable learning via sample reweighting. In *Association for the Advancement of Artificial Intelligence*, 2020. 4

[63] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019. 3

[64] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. 2

[65] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, 2017. 2

[66] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000. 5

[67] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *International Conference on Artificial Intelligence and Statistics*, 2019. 1, 3

[68] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018. 1, 2

[69] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. Preprint arXiv:2103.03097, 2021. 3

[70] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005. 3

[71] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. Preprint arXiv:1904.12848, 2019. 3

[72] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. Preprint arXiv:2106.03721, 2021. 7

[73] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. Improved ood generalization via adversarial training and pre-training. In *International Conference on Machine Learning*, 2021. 2

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3

[75] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3, 7

[76] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. Preprint arXiv:2103.02503, 2021. 2, 7

[77] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Association for the Advancement of Artificial Intelligence*, 2020. 3, 7

[78] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, 2020. 2, 3, 8

[79] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. 3

[80] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 3, 8

[81] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020. 5